Table 1 Prediction performances of four different models.

model	Accuracy	Sensitivity	Precision	MCC	AUC
N-glycosylation sites					
Window (N1)	0.767	0.494	0.658	0.412	0.814
Window + di-pep (N2)	0.884	0.766	0.840	0.721	0.942
Window + subcellular (N3)	0.822	0.640	0.743	0.568	0.891
Window + di-pep + subcellular (N4)	**0.896**	**0.808**	**0.844**	**0.752**	**0.952**
O-glycosylation sites					
Window (O1)	0.784	0.534	0.708	0.473	0.831
Window + di-pep (O2)	0.893	0.779	0.868	0.748	0.949
Window + subcellular (O3)	0.813	0.639	0.732	0.553	0.866
Window + di-pep + subcellular (O4)	**0.897**	**0.790**	**0.870**	**0.756**	**0.952**

Window means local information is used for prediction. Similarly, di-pep means the use of general information and subcellular means that of subcellular localization.

mation. Subcellular localization is determined partly by sorting signals, such as the secretory signal peptide "Ser-Lys-Leu"[25]. In fact, the frequency of a particular peptide is used to predict subcellular localization by WoLF PSORT[26]. Thus, counting the frequency of dipeptides in a protein sequence, which is used to represent general information about proteins, partly corresponds to counting signal peptides and considering subcellular localization information.

2.2 Comparison of feature representation of local information with the previous studies

Several approaches to encode local information have been proposed[14)-17)]. We compared these approaches using several lengths of the sequence window (**Fig. 2**). As shown in Fig. 2, among the BLOSUM62 profile encoding, 0/1 encoding and physico-chemical property encoding, the BLOSUM62 profile encoding system, which was used in our method, was, except when using the window of length 4, better than the other two encodings in the N-glycosylation prediction. On the other hand, in the O-glycosylation prediction (Fig. 2), the 0/1 encoding system was better than the other two encodings except when using the window of length 10. However, the difference between the performances of the 0/1 encoding system and the BLOSUM62 profile encoding system was very small. As for the window length, the prediction performances almost generally peaked when using the sequence window of length 20. Thus we adopted the BLOSUM 62 profile encoding system, using the window of length 20. Here, we confirm the superiority of our feature representation method to those used in previous studies[14)-17)]. These studies considered only local information; hence their method performance was estimated to be nearly the same as "Window" in Table 1. Therefore, as our method, utilizing the whole sequence information and subcellular localization, improved the prediction accuracy of "Window" by more than 10 percent in both N-glycosylation and O-glycosylation site prediction (Table 1), our method is competitive with and sometimes surpass the previous methods.

Table 2 O-glycosylation site prediction in three protein sequences

method	Sensitivity	Balanced accuracy
NetOGlyc[15)]	0.563	0.728
EnsembleGly[16)]	0.375	0.679
Our method	**1.000**	**0.766**

The BSP30, Kallikrein-1 and Ig delta chain C region have sixteen experimentally validated O-glycosylation sites. Previous methods (NetOGlyc and EnsembleGly) and our method, which used almost the same positive data to train the prediction model, were applied to these sites. Our method achieved 1.000 (16/16) sensitivity, while the previous methods showed 0.563 (9/16) and 0.375 (6/16) sensitivity respectively. Balanced accuracy was calculated as follows:

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}-\text{FN}} + \frac{\text{TN}}{\text{TN}-\text{FP}}\right)$$

2.3 Comparison of prediction of known O-glycosylation sites

Our method and several previous methods[15),16)] were applied to the sixteen experimentally validated O-glycosylation sites of three protein sequences, which are BSP-30, Kallikrein-1 and Ig delta chain C region (**Table 2**). Our method and the previous methods used almost the same positive data, which didn't contain BSP30, Kallikrein-1 and Ig delta chain C region, to train the prediction model. As shown in Table 2, our method achieved 1.000 (16/16) sensitivity, while the previous methods showed 0.375 (6/16) and 0.563 (9/16) sen-

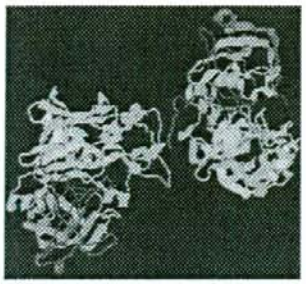

Fig. 3 3D structure of beta-secretase 1 (PDB ID:1FKN) BACE1 is an enzyme that breaks down proteins and regulates functions of membrane proteins. Furthermore, it is known to be associated with Alzheimer's disease. BACE1 forms a homo-dimer.

sitivity respectively. Moreover, our method showed better balanced accuracy than the previous methods. Hence we can conclude that in predicting O-glycosylation sites our method is competitive with and sometimes superior to the previous methods.

2.4 Validation of biological application of the proposed model to the N-glycosylation site prediction

Although the previous studies[14)–17)] focused on the O-glycosylation, our study also produced the prediction model for the N-glycosylation. To confirm the biological applicability of our prediction model, we predicted the N-glycosylation sites of a protein, envelope glycoprotein gp120 precursor, whose glycosylation sites have been identified. Gp120 was not included in the dataset.

Envelope glycoprotein gp120 precursor is a part of envelope glycoprotein from AIDA virus[27)] and has 17 consensus N-glycosylation motifs (Asn-Xaa-Ser/Thr). Among them, 14 sites are validated to be glycosylated in the PDB database (PDB ID: 1G9M).

10 out of these 14 sites were correctly predicted as glycosylation sites. Moreover, 2 of 3 non-glycosylation sites were successfully identified. Thus we conclude that our model can be applied to glycoproteins with sufficient reliability.

2.5 Predictions for unknown glycosylation sites

To validate the applicability of our prediction model at a genome-wide level, we predicted the N-glycosylation sites of beta-secretase 1 (BACE1) whose glycosylation sites have not been identified. BACE1 (**Fig. 3**) is an enzyme that breaks down proteins, and which regulates the function of membrane proteins[28)]. Moreover, it is known to be associated with Alzheimer's disease[29)].

The BACE1 protein sequence has four consensus N-glycosylation motifs (Asn-Xaa-Ser/Thr). We predicted whether these four sites would be glycosylated or not using our method (**Table 3**). Three sites were predicted to be glycosylated and the other one was predicted to be non-glycosylated. The prediction for these four sites was finished within 0.3 seconds on a 2-CPU cluster (Opteron 275 2.2 GHz processors). This fast computation suggests our method can be applied at a genome-wide level.

To confirm the validity of our predictions, the local structure around the predicted N-glycosylation sites in BACE1 as well as known N-glycosylation sites in the training dataset were shown in **Fig. 4**. The molecular mechanism of N-glycosylation is that a glycan moiety is attached to an asparagine residue by binding to the amido group in the target residue. As glycan moieties are larger than amino acids with several monosaccharides that have a ring structure, some space around the amido group of the asparagine is necessary for glycosylation to occur. In particular, the amido group of the asparagine residue shown in Fig. 4(B), a known glycosylation site, has plenty of space around it and sticks out. Similarly, the amido group of the 153rd asparagine residue predicted to be a glycosylation site, shown in Fig. 4(A), is likely to bind to a glycan moiety since there is a lot of space around it and the amido group is very exposed. On the other hand, the amido group of the 223rd asparagine residue predicted to be non-glycosylated, shown in Fig. 4(C), is less likely to be glycosylated, because the space surrounding it is as small as a known non-glycosylation site, shown in Fig. 4(D).

To assess our prediction quantitatively, we calculated the solvent-excluded surface (SES) area by MSMS[30)]. MSMS is a software which has been shown to be fast and reliable in computing molecular surfaces. The SES is the topological boundary of the union of all possible probes that do not overlap with the molecule (**Fig. 5**) and is used to visualize and study molecular properties[30)]. The SES area of each amido group of the asparagine which we predicted as glycosylation sites, 153th, 172th and

Table 3 N-glycosylation site prediction in BACE1

Residue number	Sequence window	Prediction result	SES ($Å^2$)
153	TDLVSIPHGP**N**VTVRANIAAI	Glycosylation site	26.88
172	AITESDKFFI**N**GSNWEGILGL	Glycosylation site	29.02
223	ISLYMGENVT**N**QSFRITILPQ	Non-glycosylation site	5.99
354	AITESDKFFI**N**GSNWEGILGL	Glycosylation site	24.20

The BACE1 has four consensus N-glycosylation motifs (Asn-Xaa-Ser/Thr). Among these, three sites (153^{rd}, 172^{nd} and 354^{th} residue) were predicted to be glycosylated and the other (223^{rd} residue) was predicted to be non-glycosylated. SES areas of amido group of these 3 positive sites are clearly larger than that of the negative site.

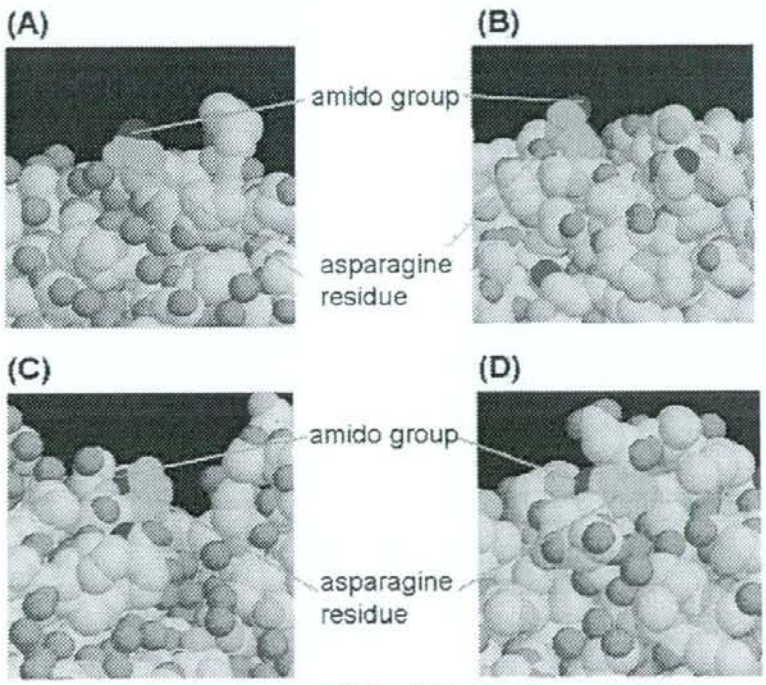

Fig. 4 Local structure around N-glycosylation sites and non-glycosylation sites. The atoms shown in green correspond to asparagine residues and atoms shown in blue illustrate an amido group in the asparagine residue. (A) The local structure around the 153^{rd} residue in BACE1, which is predicted to be a glycosylation site. (B) The local structure around a known glycosylation site in the training dataset. (C) The local structure around the 223^{rd} residue in BACE1, which is predicted to be a non-glycosylation site. (D) The local structure around a known non-glycosylation site in the training dataset.

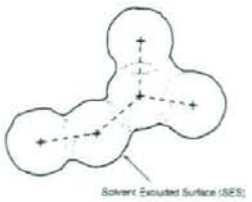

Fig. 5 The solvent-excluded surface (SES). SES is the topological boundary of the union of all possible probes having no intersection with a set of overlapping spheres M. This surface is used to not only describe hydration effects, but also to visualize protein surfaces and to study molecular properties.

354^{th} residues, is obviously larger than that of the amido group of the asparagine which we predicted as a non-glycosylation site. 223^{rd} residue (Table 3). Here, even if the molecular dynamics simulations were performed, the SES area of the amido group of the asparagine residue didn't fluctuate significantly (See Supplementary Material 3). The SES area of the glycosylated amido group was constantly larger than that of the non-glycosylated amido group.

We also applied the same evaluation approach to the O-glycosylation site prediction. O-glycosylation sites of leptin precursor which is the causal factor of adipositas were predicted[31]. The molecular mechanism of O-glycosylation is that a glycan moiety is attached to a serine or threonine residue by binding to the hydroxyl group in the target residue.

Leptin precursor has twenty-two candidate sites of O-glycosylation. Among these candidates, seven sites were predicted to be glycosylated (Supplementary Material 1).

We analyzed the local structure around the predicted O-glycosylation sites in leptin precursor (**Fig. 6**). The hydroxyl group of the 138^{th} serine residue, predicted as a glycosylation site, was shown in Fig. 6(A). On the other hand, the hydroxyl group of the 73^{rd} serine residue, predicted as a non-glycosylation site, was shown in Fig. 6(B). As shown in Fig. 6, the hydroxyl group of the 138^{th} serine was spatially more suitable for an approach of glycosyltransferases than that of the 73^{rd} serine. SES areas of the hydroxyl group in the 7 predicted glycosylation residues are significantly larger than those in the non-glycosylation residues (P-value ¡ 0.02 by t test) (See Supplementary Material 2).

Therefore, we conclude that our model can predict structurally reasonable both N- and O-glycosylation sites in proteins.

3. Discussion

Our model, which predicts glycosylation sites using not only local information, but also general information and subcellular localization of proteins, showed better prediction performances than previous models[14)-17)], which only considered local information (Table 1). These findings suggest that it is important to consider whole-protein-sequence information and subcellular localization when predicting glycosylation sites. Furthermore, in our computational experiment, in which our model was applied to a protein whose glycosylation sites had not been identified, glycosylation sites predicted by our model were shown to be structurally reasonable (Fig. 4 and Fig. 6). Therefore, we conclude that our method is a comprehensive and effective computational method that is applicable at a genome-wide level.

4. Conclusions

In the present study, we developed a comprehensive and effective computational method that detects glycosylation sites. Identification of the structure of glycans attached to glycosylation sites is a challenge that follows the identification of glycosylation sites. To resolve this problem, it is necessary to construct a comprehensive database, which contains information about glycosylation sites and glycan structures at each glycosylation site. Identification of glycosylation sites and protein-bound glycan structures will contribute to further understanding of the functions of glycosylation and glycans that have not been fully elucidated. Moreover, if we can overcome these problems, the field of glycoinformatics will be established next to bioinformatics and chemoinformatics.

5. Methods

5.1 Support vector machine

SVM is a new technique for data classification that has better performance than ANN[32)]. SVM has been used to solve a variety of biological classification problems[33)-37)].

The concept of SVM is based on the structural risk minimization principle to minimize both training and generalization errors[38)].

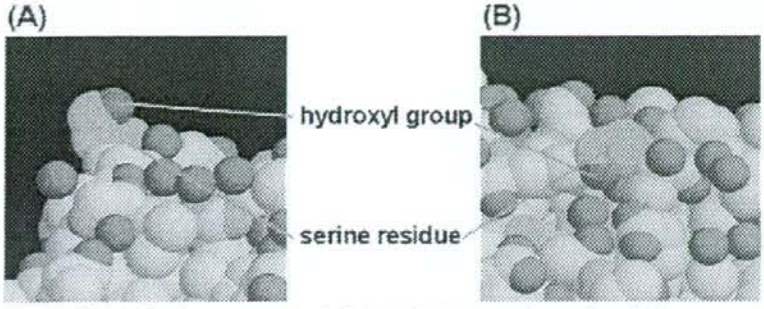

Fig. 6 Local structure around O-glycosylation sites and non-glycosylation sites. The atoms shown in green correspond to serine residues and atoms shown in purple illustrate a hydroxyl group in the serine residue. (A) The local structure around the 117th residue in leptin precursor, which is predicted to be a glycosylation site. (B) The local structure around the 52nd residue in leptin precursor, which is predicted to be a non-glycosylation site.

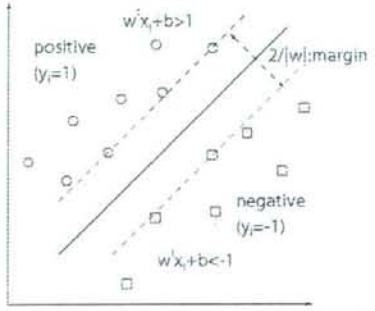

Fig. 7 Schematic diagram of SVM separating positives (circles) and negatives (squares) in a higher dimensional feature space. Hyperplanes (dotted lines) are determined so that $|w|$, the Euclidean norm of weights for each dimension or feature, is minimized, or the margin ($2/|w|$) is maximized.

When used for classification, SVM separates positive (for example, glycosylation sites) and negative (for example, non-glycosylation sites) training samples in a multidimensional space by constructing a hyperplane optimally positioned between the positive and negative samples (**Fig. 7**). A testing sample is then projected onto this multidimensional space to determine its class affiliation based on its relative position to the hyperplane.

SVM produces the classifier shown in Equation (1). In SVM, each feature vector x_i is projected into a higher dimensional feature space using a kernel function such as the RBF kernel, or $K(x_i, x_j)$ in Equation (1).

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{n} y_i \lambda_i^* K(x_i, x) - b^*\right),$$
$$K(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right) \quad (1)$$

where λ_i^* is a Lagrange multiplier, b^* is a parameter which is determined by the hyperplane and σ is a parameter of RBF Kernel.

In this paper, we used the SVM software named *LIBSVM*[39] to perform the prediction task. RBF kernel was selected as it showed the best performances (See Supplementary Material 1). Kernel functions used were as follows.

Linear kernel : $K(x_i, x_j) = x_i^T x_j$

Polynomial kernel : $K(x_i, x_j) = (\gamma x_i^T x_j)^3$

Sigmoid kernel : $K(x_i, x_j) = \tanh(\gamma x_i^T x_j)$

5.2 Extraction of a sequence descriptor

5.2.1 Local information

We encoded local information of glycosylation sites by extracting a subsequence within a window of fixed size (Fig. 1). We extracted k upstream and downstream residues of Asn (N), Ser (S) or Thr (T) residues that were predicted to be glycosylated. In this paper, we set $k = 10$, constituting the sequence window of 20 residues (Fig. 1). In case the full sequence window cannot be extracted, we define 'Z' as the 21st amino acid to fill blanks (**Fig. 8**). To encode one residue in the sequence window, we

Fig. 8 'Z' as the 21st amino acid. When the glycosylation site is near the ends of protein sequence, the full sequence window cannot be extracted. In this situation, we define 'Z' as the 21st amino acid to fill blanks.

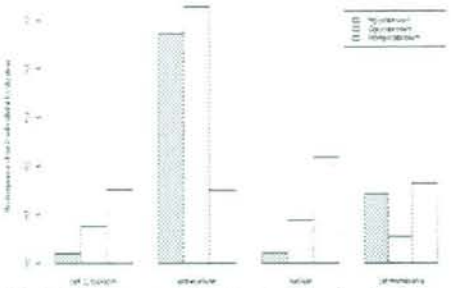

Fig. 9 The frequency of each subcellular localization. Distribution of subcellular localization prediction outputs of Wolf PSORT for glycoproteins and non-glycoproteins in our datasets is illustrated. It should be noted that the prediction output of Wolf PSORT is based on localization of proteins similar to a query and thus several localizations where N-linked glycoproteins don't exist, for example, are observed.

utilized the BLOSUM62 profile encoding (the corresponding row in the BLOSUM62 matrix). For example, the BLOSUM62 profile for alanine is equal to the vector (0.1.-1.-1.4.0.-1.-2.-1.-1.-2.-1.-1.-1.-1.-1.-2.-2.-2.-3) and that for 'Z', the 21st amino acid, is equal to the vector (0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0). Therefore, a 20×20 dimension vector was calculated for each sequence window. In the previous study[40], BLOSUM encoding, where each row in BLOSUM matrix was utilized to encode each amino acid, was used to predict T-cell class 1 epitopes by neural network. In this study, the prediction performance with this encoding was better than the other method.

5.2.2 General information about proteins

We counted the frequency of di-peptides in a whole protein sequence to encode general protein information. Glycans are attached to proteins by glycosyltransferases, which interact with the target proteins. The interaction with the objective protein depends not only on the local site but also on the whole protein structure. In order to consider the effects of glycosyltransferases, the structures of proteins should be taken into account. In a previous study, it was shown that protein structural classes can be predicted by counting the frequency of di-peptides[41]. Thus, we assume that counting the frequency of di-peptides enables consideration of protein structures. As there are 20 amino acids and 20×20 kinds of di-peptides, a 400-dimension vector was calculated for each protein.

5.2.3 Subcellular localization information about proteins

We used the output of WoLF PSORT[26] to encode subcellular localization information. Proteins are synthesized in the ribosome and modified with glycans in the endoplasmic reticulum or Golgi. The resultant glycoproteins are distributed throughout cells. In particular, most membrane proteins are glycoproteins[24]. For example, the subcellular localization of glycoproteins and non-glycoproteins in our datasets is shown in **Fig. 9**. As shown in Fig. 9, the subcellular localization of glycoproteins is specific, as about half of all glycoproteins localize extracellularly, while only 15% non-glycoproteins localize extracellularly. In WoLF PSORT, localization of the target sequence is determined based on the localization of training proteins that have sequence similarity with the target. To encode subcellular localization information, we utilized the frequency of each subcellular localization in the output of Wolf PSORT. The value for the subcellular localization x is calculated as the number of proteins localizing in x divided by the total number of proteins similar to the target. As there are 23 subcellular localizations in the output of WoLF PSORT, a 23-dimension vector was calculated for each target protein (**Fig. 10**).

5.2.4 The structure of the feature vector

To utilize all information (local information, general information and subcellular localization

Fig. 10 Encoding the output of WoLF PSORT. In this example, WoLF PSORT exhibits that, among 32 sequences similar to protein1, there are 25 proteins that localize extracellularly (*extra*). Therefore, the 3rd element of the feature vector, which corresponds to an *extra* localization, is 25/32 for protein1. The value for the subcellular localization x is calculated as the number of proteins localizing in x divided by the total number of proteins similar to the target. Here, *extr* stands for extracellular, *lyso* for lysosome, *plas* for plasmalemma, *nucl* for nuclear, *E.R.* for endoplasmic reticulum.

Fig. 11 The structure of the feature vector. The each vector (vectors derived from local information, general information or subcellular information) was combined respectively. We use the combined vector as an input for *LIBSVM*.

information), the each vector was combined respectively (**Fig. 11**). If a protein has more than one glycosylation sites, vectors derived from protein whole sequence and subcellular localization information are identical for these sites. We use the combined vector as an input for *LIBSVM*.

5.3 Prediction performance assessment

The performance of SVM has often been assessed using the five-fold cross validation method[42]. The dataset was randomly divided into five subsets of approximately equal size. One of the five subsets was used as a test set, and the remaining four subsets were used as training sets. This process was repeated five times so that every subset was used as a test set once. The performance of SVM can be assessed on the basis of accuracy, sensitivity, precision, MCC (Matthew's correlation coefficient) and AUC defined as follows:

$$\text{Accuracy} = \frac{TP - TN}{TP - TN - FP - FN}$$

$$\text{Sensitivity} = \frac{TP}{TP - FN}$$

$$\text{Precision} = \frac{TP}{TP - FP}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP - FP)(TP - FN)(TN - FP)(TN - FN)}}$$

Here, TP, TN, FP, FN stand for true positive, true negative, false positive and false negative, respectively. MCC ranges between -1 and 1. If there is no relationship between the predicted values and the real values, MCC should be around 0. In contrast, there is strong relationship between the predicted values and the real values, MCC should be close to 1. AUC represents the Area Under the (ROC) Curve which draws the evolution of the true positive rate versus the false positive rate. The AUC of an ideal classifier would be 1, while for a random classifier it would be 0.5.

5.4 Dataset construction

5.4.1 N-glycosylation site dataset

From the glycosciences.de database[43], we collected N-glycosylation sites in human proteins that were validated in PDB database[44] as positives. As putative negative data, we randomly extracted 700 Asn residues attached no N-glycosylation annotation and with a consensus motif in the form of "Asn-Xaa-Ser/Thr" (Xaa represents all kinds of amino acid except for proline) from the glycoproteins which have some annotations about glycosylation (such as "Potential", "Probable" and "By similarity") in UniProtKB/Swiss-Prot. This extraction significantly reduces the possibility to unexpectedly pick up false negatives, because in the glycoproteins with glycosylation annotation, every Asn residue site must have been examined and therefore Asn residue site with no glycosylation annotation is quite certainly true non-glycosylation site. The N-glycosylation dataset consisted of 308 positives from 125 human proteins and 700 negatives from 648 human proteins.

5.4.2 O-glycosylation site dataset

From the O-GLCBASE database[45], we collected O-glycosylation sites in mammalian proteins that were evidenced experimentally as positives. As putative negative data, we picked up the glycoproteins by choosing the proteins which have some annotations about glycosylation (such as "Potential", "Probable" and "By similarity") in UniProtKB/Swiss-Prot as mentioned above. From these limited proteins, we randomly extracted 1200 Ser/Thr residues in mammalian sequences with no annotation (such as "Potential", "Probable" and "By similarity") related to O-glycosylation in UniProtKB/Swiss-Prot. Since the mucin protein sequence has repeat sequences, several identical subsequences were generated within

the window. These identical subsequences were counted as one positive or one negative in the dataset. The O-glycosylation dataset was composed of 551 positives from 242 mammalian proteins and 1200 negatives from 1160 mammalian proteins.

These N-glycosylation and O-glycosylation site dataset are available in our web site (`http://www.dna.bio.keio.ac.jp/glycan/`).

Authors' contributions

KS developed the idea of the method, implemented the system, and executed the computational experiments and analyses. NN provided the basic idea of the study and contributed to the mathematical design and analysis of the method. YS participated in the design and coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by a Grant program for bioinformatics research and development from the Japan Science and Technology Agency, and a Grant-in-Aid for Scientific Research on Priority Area No. 17018029.

Supplementary Materials

(1) Comparison with the performances by using other kernels.
Integral model was utilized and different kernels (RBF, linear, polynomial and sigmoid) were applied in SVM computation.

(2) O-glycosylation site prediction in leptin precursor.
Leptin precursor has twenty two candidate sites of O-glycosylation. Sequence windows around the candidate sites and the SES area of hydroxyl group are shown as well as the prediction result.

(3) Effect of conformational change on SES area.
The average SES area of both the glycosylated and non-glycosylated amido group in several conformation of the endothelial protein C receptor precursor (PDB ID: 1L8J) is shown. One nanosecond molecular dynamics simulation was performed with AMBER 9[46], and the SES area was calculated every 200 picoseconds.

References

1) Salas, J. and Mendez, C.: Engineering the glycosylation of natural products in actinomycetes, *Trends Microbiol.*, Vol.15, pp.219–232 (2007).
2) Saxon, E. and Bertozzi, C.: Chemical and biological strategies for engineering cell surface glycosylation, *Annu. Rev. Cell Dev. Biol.*, Vol.17, pp.1–23 (2001).
3) Plante, O.: Combinatorial chemistry in glycobiology, *Comb. Chem. High Throughput Screen.*, Vol.8, pp.153–159 (2005).
4) Breton, C., Snajdrova, L., Jeanneau, C., Koca, J. and Imberty, A.: Structures and mechanisms of glycosyltransferases, *Glycobiology*, Vol.16, pp.29R–37R (2006).
5) Breton, C. and Imberty, A.: Structure/function studies of glycosyltransferases, *Curr. Opin. Struct. Biol.*, Vol.9, pp.563–571 (1999).
6) Imberty, A., Wimmerova, M., Koca, J. and Breton, C.: Molecular modeling of glycosyltransferases, *Methods Mol. Biol.*, Vol.347, pp.145–156 (2006).
7) Goletz, S., Thiede, B., Hanisch, F., Schultz, M., Peter-Katalinic, J., Muller, S., Seitz, O. and Karsten, U.: A sequencing strategy for the localization of O-glycosylation sites of MUC1 tandem repeats by PSD-MALDI mass spectrometry, *Glycobiology*, Vol.7, pp.881–896 (1997).
8) Sadeghi, H. and Birnbaumer, M.: O-Glycosylation of the V2 vasopressin receptor, *Glycobiology*, Vol.9, pp.731–737 (1999).
9) Skropeta, D., Settasatian, C., McMahon, M., Shearston, K., Caiazza, D., McGrath, K., Jin, W., Rader, D., Barter, P. and Rye, K.: N-Glycosylation regulates endothelial lipase-mediated phospholipid hydrolysis in apoE- and apoA-I-containing high density lipoproteins, *J. Lipid Res.*, Vol.48, pp.2047–2057 (2007).
10) Wojczyk, B., Takahashi, N., Levy, M., Andrews, D., Abrams, W., Wunner, W. and Spitalnik, S.: N-glycosylation at one rabies virus glycoprotein sequon influences N-glycan processing at a distant sequon on the same molecule, *Glycobiology*, Vol. 15, pp. 655–666 (2005).
11) Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K., Ueda, N., Hamajima, M., Kawasaki, T. and Kanehisa, M.: KEGG as a glycome informatics resource, *Glycobiology*, Vol.16, pp.63R–70R (2006).
12) Jenkins, N., Parekh, R. and James, D.: Getting the glycosylation right: implications for the biotechnology industry, *Nat. Biotechnol.*, Vol.14, pp.975–981 (1996).

13) Apweiler, R., Hermjakob, H. and Sharon, N.: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database, *Biochim. Biophys. Acta*, Vol.1473, pp.4–8 (1999).
14) Li, S., Liu, B., Zeng, R., Cai, Y. and Li, Y.: Predicting O-glycosylation sites in mammalian proteins by using SVMs, *Comput Biol Chem*, Vol.30, pp.203–208 (2006).
15) Julenius, K., Molgaard, A., Gupta, R. and Brunak, S.: Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites, *Glycobiology*, Vol.15, pp.153–164 (2005).
16) Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D. and Honavar, V.: Glycosylation site prediction using ensembles of Support Vector Machine classifiers, *BMC Bioinformatics*, Vol.8, p.438 (2007).
17) Chen, Y., Tang, Y., Sheng, Z. and Zhang, Z.: Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, *BMC Bioinformatics*, Vol.9, p.101 (2008).
18) Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. and Brunak, S.: Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence, *Proteomics*, Vol. 4, pp. 1633–1649 (2004).
19) Gupta, R. and Brunak, S.: Prediction of glycosylation across the human proteome and the correlation to protein function, *Pac Symp Biocomput*, pp.310–322 (2002).
20) Petrescu, A., Milac, A., Petrescu, S., Dwek, R. and Wormald, M.: Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding, *Glycobiology*, Vol.14, pp.103–114 (2004).
21) Baenziger, J.: Protein-specific glycosyltransferases: how and why they do it!, *FASEB J.*, Vol.8, pp.1019–1025 (1994).
22) Opdenakker, G., Rudd, P., Ponting, C. and Dwek, R.: Concepts and principles of glycobiology, *FASEB J.*, Vol.7, pp.1330–1337 (1993).
23) vonder Lieth, C., Bohne-Lang, A., Lohmann, K. and Frank, M.: Bioinformatics for glycomics: status, methods, requirements and perspectives, *Brief. Bioinformatics*, Vol.5, pp.164–178 (2004).
24) Spiro, R.: Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds, *Glycobiology*, Vol.12, pp.43R–56R (2002).
25) Nielsen, H., Brunak, S. and von Heijne, G.: Machine learning approaches for the prediction of signal peptides and other protein sorting signals, *Protein Eng.*, Vol.12, pp.3–9 (1999).
26) Horton, P., Park, K., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. and Nakai, K.: WoLF PSORT: protein localization predictor, *Nucleic Acids Res.*, Vol.35, pp.W585–587 (2007).
27) Ohgimoto, S., Shioda, T., Mori, K., Nakayama, E., Hu, H. and Nagai, Y.: Location-specific, unequal contribution of the N glycans in simian immunodeficiency virus gp120 to viral infectivity and removal of multiple glycans without disturbing infectivity, *J. Virol.*, Vol.72, pp.8365–8370 (1998).
28) Zacchetti, D., Chieregatti, E., Bettegazzi, B., Mihailovich, M., Sousa, V., Grohovaz, F. and Meldolesi, J.: BACE1 expression and activity: relevance in Alzheimer's disease, *Neurodegener Dis*, Vol.4, pp.117–126 (2007).
29) Heneka, M. and O'Banion, M.: Inflammatory processes in Alzheimer's disease, *J. Neuroimmunol.*, Vol.184, pp.69–91 (2007).
30) Sanner, M., Olson, A. and Spehner, J.: Reduced surface: an efficient way to compute molecular surfaces, *Biopolymers*, Vol.38, pp. 305–320 (1996).
31) Sone, M. and Osamura, R.: Leptin and the pituitary, *Pituitary*, Vol.4, pp.15–23 (2001).
32) Byvatov, E. and Schneider, G.: Support vector machine applications in bioinformatics, *Appl. Bioinformatics*, Vol.2, pp.67–77 (2003).
33) Bhasin, M. and Raghava, G.: GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors, *Nucleic Acids Res.*, Vol.32, pp.W383–389 (2004).
34) Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. and Suwa, M.: GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model, *Nucleic Acids Res.*, Vol.33, pp.W148–153 (2005).
35) Gubbi, J., Shilton, A., Parker, M. and Palaniswami, M.: Protein topology classification using two-stage support vector machines, *Genome Inform*, Vol.17, pp.259–269 (2006).
36) Han, L., Zheng, C., Xie, B., Jia, J., Ma, X., Zhu, F., Lin, H., Chen, X. and Chen, Y.: Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness, *Drug Discov. Today*, Vol.12, pp.304–313 (2007).
37) Burbidge, R., Trotter, M., Buxton, B. and Holden, S.: Drug design by machine learning: support vector machines for pharmaceutical data analysis, *Comput. Chem.*, Vol.26, pp.5–14 (2001).
38) Baldi, P., Brunak, S., Chauvin, Y., Ander-

sen, C. and Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics*, Vol.16, pp.412–424 (2000).
39) Chang, C.-C. and Lin, C.-J.: *LIBSVM: a library for support vector machines* (2001). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
40) Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S., Lamberth, K., Buus, S., Brunak, S. and Lund, O.: Reliable prediction of T-cell epitopes using neural networks with novel sequence representations, *Protein Sci.*, Vol.12, pp.1007–1017 (2003).
41) Chen, C., Zhou, X., Tian, Y., Zou, X. and Cai, P.: Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, *Anal. Biochem.*, Vol.357, pp.116–121 (2006).
42) Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y. and Chen, Y.: Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity, *Proteomics*, Vol.6, pp.4023–4037 (2006).
43) Lutteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M. and vonder Lieth, C.: GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research, *Glycobiology*, Vol.16, pp.71R–81R (2006).
44) Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P.: The Protein Data Bank, *Nucleic Acids Res.*, Vol.28, pp.235–242 (2000).
45) Gupta, R., Birch, H., Rapacki, K., Brunak, S. and Hansen, J.: O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins, *Nucleic Acids Res.*, Vol.27, pp.370–372 (1999).
46) Case, D., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz, K., Onufriev, A., Simmerling, C., Wang, B. and Woods, R.: The Amber biomolecular simulation programs, *J Comput Chem*, Vol.26, pp.1668–1688 (2005).

Current Trends in Mecinal Chemistry, in press, 2009.

Current Trends in Medicinal Chemistry

Review

Synthesis of bio-active compounds from cyclitol derivatives provided by bioconversion of *myo*-inositol

Seiichiro Ogawa* and Miki Kanto

Department of Biosciences and Informatics, Faculty of Science and Technology, Keio University, Hiyoshi, Kohoku-ku, Yokohama, 223-8522 Japan

ABSTRACT

Biogenesis of *myo*-inositol has been shown to readily produce adequate amounts of optically active deoxyinositols, (+)-*epi*- and (–)-*vibo*-quercitols, which can be applied as versatile synthetic intermediates for development of new cyclitol derivatives of biological interest. In this review we describe new preparation of 1,3,4-trisphosphates of 3- and 6-deoxy-*myo*-inoisitols, several 1,2- and 2,3-anhydro-6-deoxyinositols, including potent glycosidase inhibitors, and biologically active deoxyinosamines. The synthetic approach adopted provides a basis for further design and synthesis of bioactive cyclitol derivatives.

KEYWORDS: inositols, quercitols, deoxy-*myo*-inositol trisphosphates, anhydrodeoxyinositols, deoxyinosamines

INTRODUCTION

In a preceding article [1] we described synthesis of biologically important branched-chain cyclitol derivatives, (–)-β-valiol and (–)-valiolamine, starting from (–)-*vibo*-quercitol (**5**), one of three deoxyinositols (quercitols) obtained through biogenesis [2] of *myo*-inositol (**1**). The synthetic route established a link between naturally abundant *myo*-inositol and chiral carbasugars [1a, 3], generally applicable for provision of large quantities of desired aminocyclitols of biological interest. In this article, we review convenient preparative routes using two quercitols **4** and **5** for several biologically active cyclitol derivatives other than carbasugar analogues (Fig. 1).

Selective blocking of hydroxyl groups of quercitols is certainly an important initial step to provide effective precursors for chemical modification leading to target compounds. Acetalation of cyclitols is a reliable way to protect pairs of vicinal hydroxyls in both cis and trans configurations: five hydroxyl groups of quercitols readily form di-*O*-isopropylidene derivatives under conventional acetalation conditions, giving one hydroxyl group unprotected derivatives.

First, synthesis of biologically important 1,4,5-trisphosphates of deoxyinositols was carried out by conventional phosphorylation of OH unprotected compounds derived from protected quercitols [4]. Secondly, intramolecular nucleophilic reaction of quercitol tosylates was conducted in order to generate anhydrodeoxyinositols, some of which have activity as glycosidase inhibitors [5]. Thirdly, three protected deoxyinososes obtained by oxidation of OH unprotected derivatives were subjected to common electrophilic reactions applied for carbonyl functions, leading to establishment of new branched-chain aminocyclitols [6]. Finally, nucleophilic substitution of quercitol tosylates with azide anions was conducted to study reaction courses controlled by stereochemistry of parent cyclitols, affording deoxyinosamines of biological interest [7].

*Corresponding author
sogawa379@ybb.ne.jp

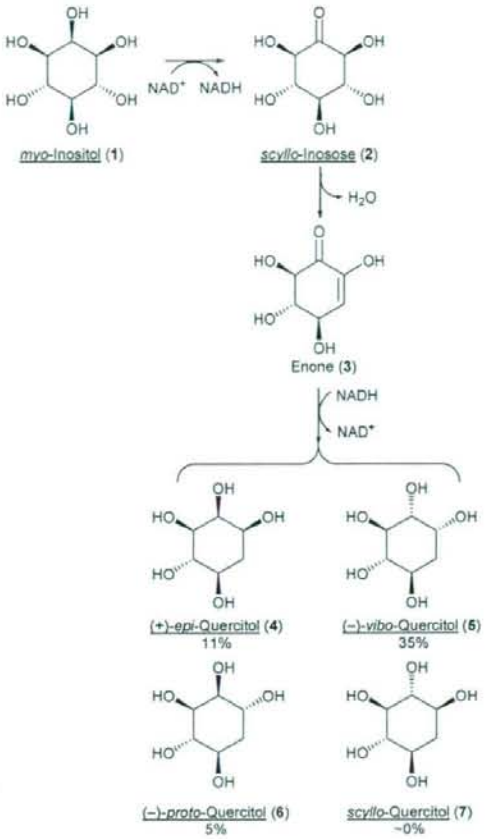

Fig. 1. Biogenesis of *myo*-inositol by *Salmonella typhimurium*. Production of three quercitols (deoxyinositols).

1. Biogenesis of *myo*-Inositol by *Salmonella typhimurium*: Production of some quercitols (deoxyinositols)

myo-Inositol [8] (**1**) is the most abundant cyclitol occurring in nature. Synthetic studies on inositol derivatives have often been complicated by difficulty in obtaining the optically active compounds desired. When *myo*-inositol is chosen as the starting material, chemical modification and/or substitution of one of the hydroxyl groups at C-1, 3, 4, and 6 leads to racemic compounds. The bioconversion of inositols therefore offers a very advantageous route to provide optically pure raw materials for cyclitol synthesis.

Recently, Takahashi *et al.* [2] succeeded in the generation of three optically active quercitols by biotransfomation of *myo*-inositol using several strains of *Salmonella typhimurium* (Fig. 1). The quercitols were isolated pure from fermentation broth by a combination of ion exchange chromatography and subsequent crystallization, the major products, being (–)-*vibo*-quercitol (**5**, 35%), followed by (+)-*epi* (**4**, 11%) and (–)-*proto*-quercitol (**6**, 5%). The mechanism of their biotransformation may be proposed as an initial bio-oxidation of *myo*-inositol (**1**) to *scyllo*-inosose (**2**), followed by dehydration and reduction, as observed by Angyal *et al.* [9] in their studies on the chemical behavior of *epi*- and *scyllo*-inososes (**2**) in neutral and/or aqueous sodium carbonate solutions. Conversion of **2** into unsaturated ketones via dehydration was observed to form an equilibrium mixture of enol-ketone (**3**) and isomers (Fig. 2). The mechanism of transformation

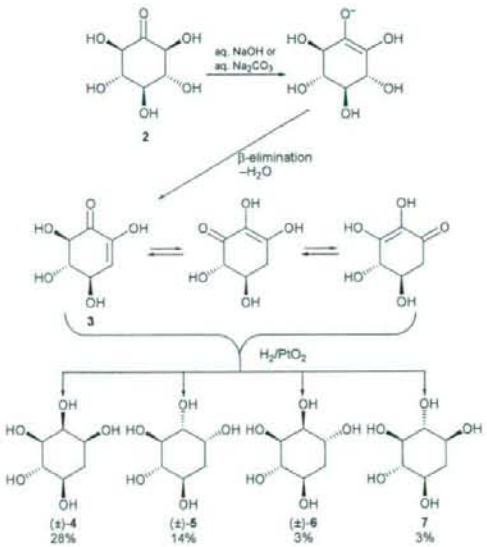

Fig. 2. *scyllo*-Inosose in alkaline solution (aq. NaOH or aq. Na_2CO_3). Formation of four quercitols, together with 12 products (cyclohexanetetrols and triols), was observed. In addition, hydrogenation over Raney nickel produced four cyclohexanepentols, including *epi*-quercitol as a major product (ca. 40% yield). All compounds are racemic and the formulae depict only one of the respective enantiomers.

of *scyllo*-inosose could be confirmed by determination of the structures of the products. Thus, catalytic hydrogenation of the reaction mixture produced all theoretically possible deoxyinositols derived from enol-ketones: DL-*epi*-quercitol (**4**, 28%) was shown to be the major product, along with DL-*vibo* (**5**, 14%), DL-*proto* (**6**, 3%), and *scyllo*-quercitols (**7**, 3%), verifying the postulated mechanism of chemical conversion of *scyllo*-inosose in an aqueous alkaline solution. In addition, hydrogenation in the presence of Raney nickel produced four cyclohexanepentols, including racemic **4** as a major product in ca. 40% yield.

2. Preparation of useful synthetic precursors: Acetalation of quercitols

Protection of the hydroxyl groups of cyclitols could be a general initial step for design of synthetic routes to target compounds. It is thus important to explore conventional protection of individual stereoisomers of cyclitols with acetal or acyl groups, considering stereochemical reaction courses and the expected reactivity of unprotected hydroxyl groups. We should always pay careful attention to possible facile migration of acetal protecting groups to neighboring free hydroxyls under acidic and/or basic conditions. Further chemical transformation of acetal-derivatives must therefore be conducted in basic media.

O-Isopropylidenation [5, 6] of (+)-*epi*-quercitol (**4**), 1D-1,2,3,5/4-cyclohexanepentol, was conducted with an excess of 2,2-dimethoxypropane (DMP, 5 molar equiv.) in DMF in the presence of TsOH (0.1 molar equiv.), the progress of the reaction being monitored by TLC (Fig. 3). When the reaction reached equilibrium, the mixture was neutralized with Na_2CO_3 and the products were separated on a silica gel column to give all three possible di-*O*-isopropylidene derivatives **8a** (26%), **9a** (24%), and **10a** (31%), whose structures were verified by treatment with *p*-TsCl in pyridine, giving the corresponding tosylates **8b–10b**.

Similar treatment of (–)-*vibo*-quercitol (**5**), 1L-1,2,4/3,5-cyclohexanepentol, with DMP in DMF gave an inseparable mixture (86%) of two di-*O*-isopropylidene derivatives [5] **11a** and **12a**. When the mixture was tosylated, the resulting compounds were easily separable by a silica gel column to give the tosylates **11b** (56%) and **12b** (43%). On

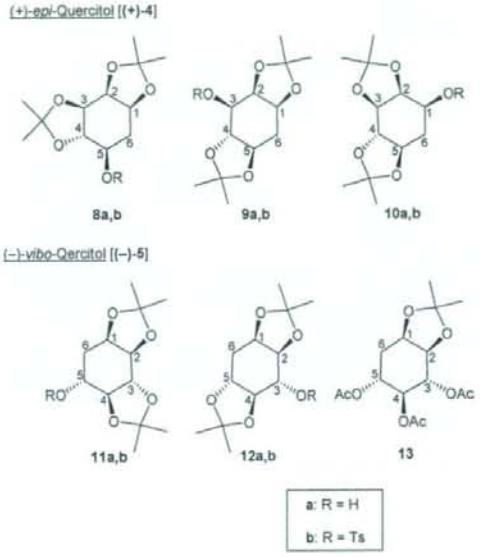

Fig. 3. Some *O*-isopropylidene derivatives and their tosylates derived from (+)-*epi*- and (–)-*vibo*-quercitols.

the other hand, the same mixture directly subjected to partial de-*O*-isopropylidenation by treatment with trace *p*-TsOH in MeOH, followed by acetylation with Ac_2O/Pyr, gave rise to 1,2-*O*-isopropylidene triacetate (**13**) (80%).

3. Synthesis of 3- and 6-Deoxy-*myo*-inositol Trisphosphates

In recent years, D-*myo*-inositol-1,4,5-trisphosphate [**14**, Ins(1,4,5)P_3], as well as its bis and tetrakisphosphates, have been demonstrated to play important roles as secondary messengers controlling many cellular processes by generating internal calcium signals, which then diffuse through the cytosol and bind to receptors on the endoplasmic reticulum causing release of calcium ions (Ca^{2+}) into the cytosol (Fig. 4). Therefore, it is feasible that inhibitors of enzymes of the phosphoinositide cascade could be of medicinal interest and also invaluable tools to elucidate the individual roles of metabolites in the regulation of cell function. In order to study biochemical and medicinal properties of polyphosphates, a large number of analogues and derivatives have so far been synthesized [10] and tested for biological

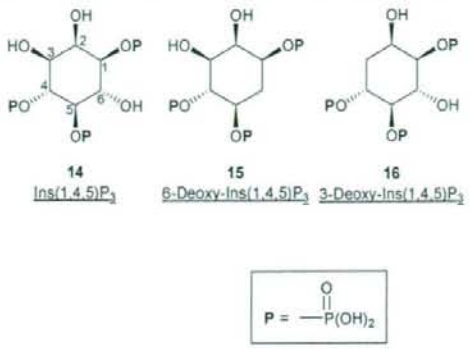

Fig. 4. *myo*-Inositol 1,4,5-trisphosphate (**14**) and its related deoxy derivatives **15** and **16**, of biological interest.

activity. Recent findings of insulin-like and anti-inflammatory properties have also stimulated us to develop means for routine synthesis of such compounds.

In this section, we describe convenient methods for a number of trisphosphate derivatives **15** and **16** of 6- and 3-deoxy-D-*myo*-inositols. Recently, synthesis of polyphosphate derivatives of 6-deoxy-D-*myo*-inositol (**4**) has been elaborated [11] from precursors derived from D-galactose, and their biological activity assayed. 6-Deoxy Ins(1,4,5)P_3 is recognized by the highly selective 3-kinase and the kinetics of its metabolism indicate that it is a substrate with resultant competitive inhibition of phosphorylation of Ins(1,4,5)P_3.

Di-*O*-isopropylidene derivatives **8a** and **10a** could be partially de-*O*-isopropylidenated with TsOH in EtOH at 0°C to give the triols **17** (70%) and **20** (78%), respectively [4] (Fig. 5). Possible contamination of these compounds due to acid-catalyzed migration of *cis*-isopropylidene groups was not observed. Compound **20** was phosphorylated to give the protected precursor **21** (60%) of 6-deoxy Ins(1,4,5)P_3 (**15**). The structure of **21** was indirectly confirmed with reference to the ^{1}H NMR spectrum of isomeric trisphosphate **19** obtained for reference by phosphorylation (→ **18**) of **17** followed by deprotection.

A mixture of **11a** and **12a** was treated with NaH in DMF and then with an excess of BnBr to give benzyl ethers, which were partially de-*O*-isopropylidenated under the influence of CSA in

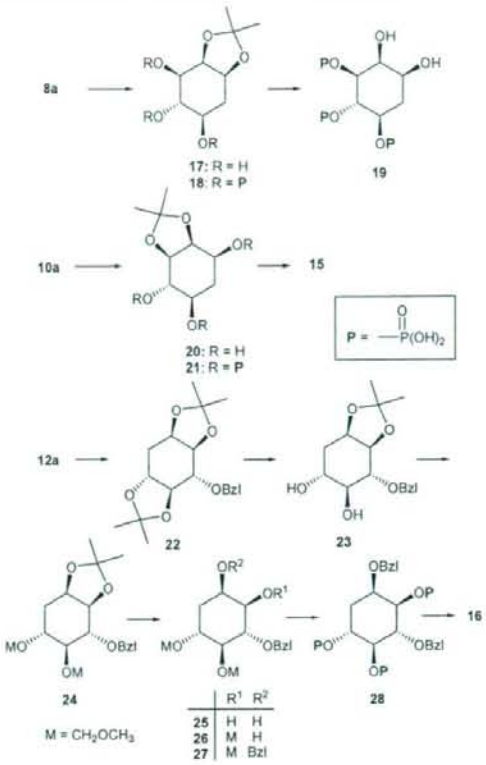

Fig. 5. Synthesis of some deoxyinositol trisphosphates.

MeOH to afford, after separation over a silica gel column, the desired 6-*O*-benzyl derivative **22** (55%). Compound **23**, obtained by partial de-*O*-isopropylidenation of **22**, was treated with $MeOCH_2Cl$ (4 molar equiv.) and diisopropylethylamine to give the di-*O*-methoxymethyl derivative **24** (89%), de-*O*-isopropylidenation of which with 80% aqueous acetic acid gave the diol **25** (88%). Treatment of **25** with dibutyltin oxide and tetrabutyl ammonium bromide, and subsequent similar etherification, gave crude methoxymethyl ether **26** (87%). In addition **26** was conventionally benzylated to give the 2-*O*-benzyl derivative **27** (91%). The methoxymethyl groups of **27** were removed by treatment with 4 M hydrochloric acid, and the product was subsequently acetylated to give the tri-*O*-acetyl derivative. This was treated with methanolic NaOMe under Zemplén conditions,

and the resulting triol was phosphorylated under the influence of dibenzyl diisopropylphopshoro-amidite (6 molar equiv.) in DMF at room temperature, and, then the reaction mixture was further treated with *m*CPBA (10 molar equiv.). The product was isolated by chromatography on silica gel to afford the 1,4,5-tris(dibenzylphosphate) **28** (93% overall yield). Hydrogenolysis of **28** in the presence of 10% Pd/C in aqueous EtOH under an atmospheric pressure of hydrogen at room temperature gave the trisphosphate, treatment of which with cyclohexylamine produced a crystalline amine salt. This was deaminated by passage through a column of Dowex 50 × 2 resin (H^+) resin to afford the free phosphate isolated as a bis-sodium salt **16** (97%).

The trisphosphates **15** and **16** did not activate pyruvate dehydrogenase phosphatase (PDH-Pase), or inhibit pyruvate dehydrogenase kinase (PDH-K) significantly. None of the compounds tested inhibited glucose 6-phosphatase (G6Pase) significantly.

4. Synthesis of anhydrodeoxyinositols

In 1971 Kupchan described [12] isolation of a naturally occurring acylated dianhydro-*C*-(hydroxymethyl)inositol, crotepoxide (**29**), shown to be interesting anticancer reagent. Cyclophellitol [13] (**30**) and the conduritol B epoxide [14] (**31**), 1,2-anhydro-L-*myo*-inositol, are known to be potent and specific inhibitors of glucocerebrosidase (Fig. 6), explained [15, 16] on the bases both of their structural resemblance to the D-glucopyranosyl cation probably formed during hydrolysis of glucosides and of covalent bonding to the active site of the enzyme through nucleophilic cleavage of the epoxide ring by the carboxylate function of aspartate or glutamate residue. Therefore, determination of inhibitory activity of the corresponding dehydroxymethyl or 3-deoxy derivative **L-55a** might be very important for understanding of structure activity relationships of inhibitors of this type. Therefore, attempts have been made to furnish several optically active 1,2- and 2,3-anhydro-6-deoxyinositols of biological interest, utilizing designed tosylates.

Removal of two isopropylidene groups of **8b** with 80% aq AcOH gave the tosylate **32**, which was treated with NaOMe (1.5 molar equiv.) in MeOH

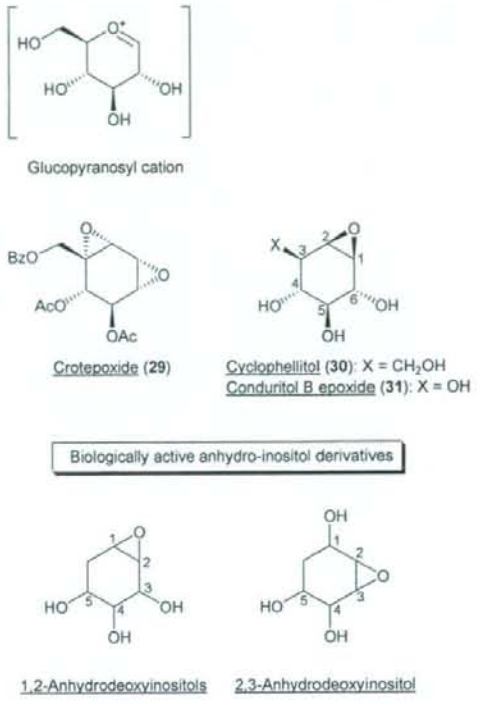

Fig. 6. Synthesis of 1,2- and 2,3-anhydrodeoxyinositols of biological interest.

at room temperature (Fig. 7). The resulting major anhydride **D-39** was obtained by use of a silica gel column in 65% yield. Similarly, **33** derived from **9b** could be converted into **L-40** (59%). The proposed structures of **D-39a** and **L-40a** were assigned on the bases of the reaction sequence and their ^{1}H NMR spectra.

The trans 4,5-*O*-isopropylidene group of **12b** was selectively removed under controlled acidic conditions to give, after acetylation, **36** (77%), a similar treatment of which with NaOMe/MeOH gave **43** (69%) and **44** (11%). These were shown to be interconvertible through epoxide group-migration under these conditions. The initially formed 3,4-anhydride **43** is likely to be attacked by a *trans*-situated 5-hydroxyl group to give 4,5-anhydride **44** and the product-ratio of these anhydrides at equilibrium would reflect their relative stability, i.e. thermodynamical features under the basic conditions. Treatment of **43** and

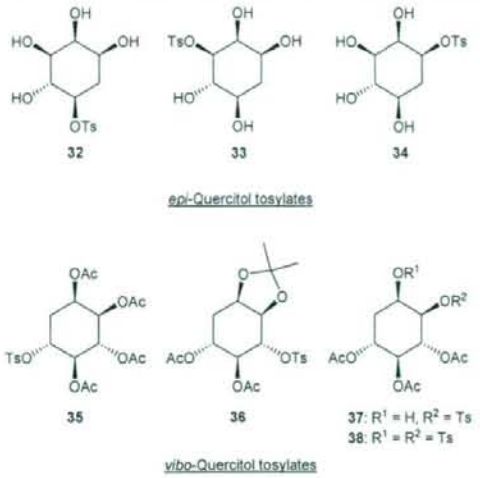

Fig. 7. Seven readily available free and protected quercitol tosylates.

44 with 50% aq AcOH gave **L-45a** (82%) and **L-39a** (68%), respectively (Fig. 8).

The 5-*O*-tosylate **35** obtained from **11b** was similarly treated with NaOMe/MeOH, and the products were acetylated to give the triacetyl derivative **D-41b** (34%) and a ca. 3:1 mixture (28%) of the triacetates **D-40b** and **L-42b**. Zemplén de-*O*-acetylation of **41b** gave **D-41a** (74%). De-*O*-acetylation of the mixture afforded, after chromatography, **D-40a** (14%) and 1L-1,2,3,5/4-cyclohyxanepentols (**L-42a**, 7%). The structures of three anhydrides formed from **35** were first deduced by considering epoxide group-migration, and assigned on the basis of the ^{1}H NMR spectra.

De-*O*-isopropylidenation of **13** with aq AcOH gave the diol, selective tosylation (1.5 molar equiv. TsCl/Pyr) of which gave the 2-tosylate **37** (93%), along with the 1,2-ditosylate **38** (7%). Similar base-treatment of **37** followed by acetylation gave the triacetate **L-46b** (85%), which afforded **L-46a** (~100%). Treatment of **38** with NaOMe/MeOH under kinetic control gave, after acetylation, the epoxide **47**, which was hydrolyzed with 10% H_2SO_4/aq acetone, followed by acetylation, giving a sole tetraacetate **48**. Similar epoxidation of **48** gave the triacetate **D-49b** (72%) (Fig. 9), which afforded **D-49a** (65%).

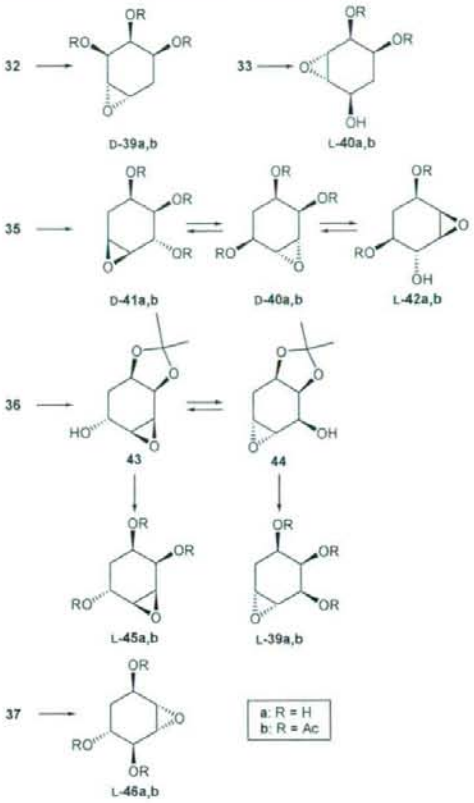

Fig. 8. Synthesis of 1,2- and 2,3-anhydrocyclohexanepentols.

Alternatively, selective benzoylation of quercitols **4** and **5** afforded directly **50** (68%) and **53** (60%), respectively. The axially oriented hydroxyl groups could hardly be esterified. Treatment of the two products with a slight excess of SO_2Cl_2/Pry gave the respective chlorides **51** (93%) and **54** (95%) with inversion of the configurations. Similar treatment of **51** under kinetic control followed by acetylation afforded two epoxides, **D-42b** (17%) and **D-52b** (43%). On the other hand, **54** afforded a sole anhydride **L-55b** (45%). De-*O*-acetylation of the triacetates gave the free anhydrides **D-42a**, **D-52a**, and **L-55a** quantitatively.

All epoxides were tested for inhibitory activity against glucocerebrosidase (mouse liver) and galactocerebrosidase (mouse liver). Among twelve

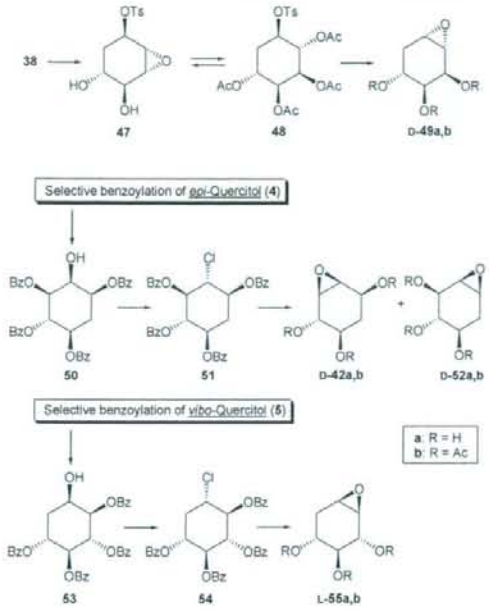

Fig. 9. Synthesis of anhydrodeoxyinositols, in addition, through selective benzoylation, chlorination, and subsequent base treatment.

stereoisomers synthesized, **L-55a** proved to be a highly potent and specific inhibitor (IC_{50} = 0.96 μM) of glucocerebrosidase, comparing favorably with **31** (IC_{50} = 8.9 μM). Other epoxides did not show inhibitory activity at $<10^{-4}$ M. Interestingly, **L-55a** did not show any inhibitory activity against galactocerebrosidase. With the three deoxy derivatives **L-42a**, **D-52a**, and **L-55a** of **30b**, only the 3-deoxy one **L-55a** was found to possess inhibitory activity comparable to **31**. Therefore, its contiguous three hydroxyl functions at C-3, -4, and -5 appear to be very important, correlating with those at C-2, -3, and -4 of the D-glucopyranosyl cation, respectively. Furthermore, the 4-, 5-, and 6-hydroxyl groups of **5** seem to be indispensable for the epoxide group to suffer nucleophilic attack against the carboxylate function of the enzyme. Interestingly, the positional isomer **L-46a** of **L-55a** was shown to be a moderate inhibitor (IC_{50} = 98 μM) of glucocerebrosidase, suggesting that the 1L-(1,2,4/3)-1,2-anhydrocyclohexanetetrol core structure correlates with the glucopyanosyl cation.

5. Synthesis of protected deoxyinososes and application to electrophilic reactions: Synthesis of aminomethyl-branched quercitols

The best known of the inososes (pentahydroxy-cyclohexanones) are *scyllo*-inosose (*myo*-inosose-2, **2**) and DL-*epi*-inosose (DL-*epi*-inosose-2, **57**), obtained from *myo*-inositol by moderate oxidation with nitric acid, and by oxidation with *Acetobacter* or by catalytic aerial oxidation, respectively.

Reactions of the carbonyl group of inososes include the addition of diazoalkanes, dithioacetal formation, reduction, hydrogenolysis, and phenylhydrazone formation. The spiro-epoxide, which is formed from *scyllo*-inosose penta-acetate and diazomethane, is the starting material for a considerable series of seven-carbon derivatives [8c].

Direct protection of inososes by acid-catalysed acylation often results in elimination of acyloxyl groups, giving isomeric enones. Direct acetalation of inososes is usually accompanied by hydration affecting keto functions, producing triacetals. Acetonation of *scyllo*- **2** and *epi*-inososes **57**, for example, has been shown to give chemically stable tri-*O*-isopropylidene derivatives [17] **56**, and **58** and/or **59**, respectively, with undesirable masking of free keto functions (Fig. 10).

In our laboratory we have prepared isomeric inosose derivatives by oxidation of protected quercitols **8a–10a** and examined their reactivity to

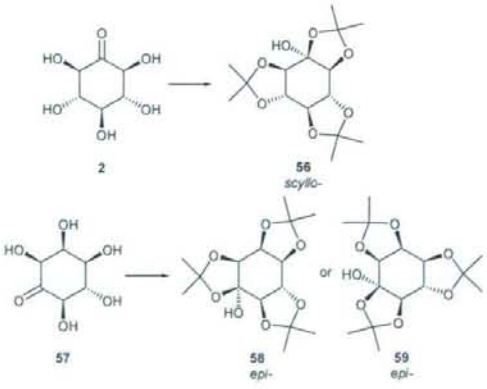

Fig. 10. Tri-*O*-isopropylidene derivatives of the hydrates generated from *scyllo*- and *epi*-inososes.

electrophilic reactions. Oxidation of **8a**–**10a** with Ac_2O/DMSO gave rise to the respective ketones **60** (92%), **61** (70%), and **62** (96%), respectively, shown to exist in keto forms and expected to be reactive synthetic intermediates for a wide variety of deoxyinositol derivatives (Fig. 11).

In an attempt to obtain exo-methylene derivatives, the ketones were first subjected to the Wittig reaction with bromotriphenylmethane in the presence of NaH-MDS, but this was not successful. Next, a base-catalyzed aldol condensation **60**–**62** was investigated [6] using an excess of nitromethane. The reaction proceeded selectively to give moderate yields of nitromethyl-branched derivatives **63a**–**65a** as single isomers. The reaction proceeded very slowly, being largely influenced by the type of base catalyst. Compound **62** readily reacted in the presence of NaOMe in MeOH to give condensate **65a**, but difficulties were encountered with **60** and **61**. These only reacted in 1 M aqueous sodium hydroxide solution, giving **63** and **64**. Considering

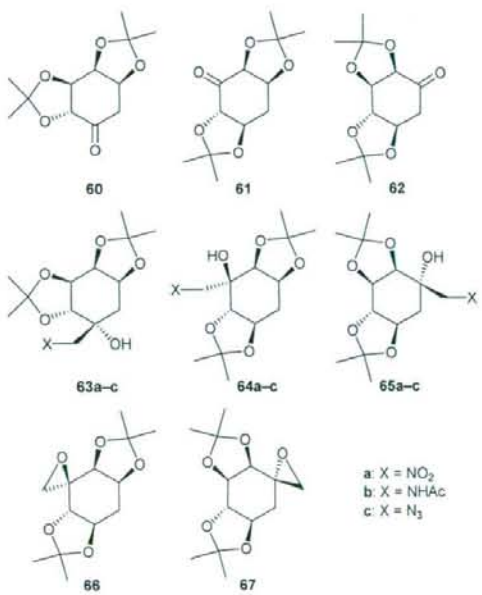

Fig. 11. Preparation of new protected deoxyinososes and their reactivity toward electrophiles, nitromethane and diazomethane.

the ^{1}H NMR spectral data, the ketones **60**–**62** adopt somewhat distorted chair-conformations. The selectivity of the aldol reaction seemed to be controlled by the steric hindrance exerted by 1,3-diaxial protons, rather than by the bulky isopropylidene groups that point away from the cyclose carbonyl group. The nitro compounds **63**–**65** were readily hydrogenated in EtOH containing Ac_2O in the presence of the Raney nickel catalyst, being converted into the respective *N*-acetyl derivatives **63b**–**65b**.

In the next reaction sequence, construction of spiro epoxides was accomplished by exposing the ketones **60**–**62** to CH_2N_2 in Et_2O–DMSO. Two spiro epoxides **66** and **67** were obtained selectively in moderate yields from **61** and **62**. Cleavage of the oxirane ring with an azide ion proceeded smoothly giving rise to the azidomethyl compounds **64c** (50%) and **65c** (65%), respectively, the structures of which were established on the basis of their ^{1}H NMR spectra, and also by comparison with those of corresponding **64a** and **65a**. The structures of **66** and **67** are shown in Fig. 11. Diazomethane was added to the ketones **61** and **62** in a similar fashion as observed in the aldol reaction. Attempts were not made to isolate the side-products, including ring-expansion products [18] likely to be formed in these reactions.

De-*O*-isopropylidenation of **63b**–**65b** with aqueous acetic acid, followed by conventional acetylation with Ac_2O in pyridine, gave the corresponding hexa-*N*,*O*-acetyl derivatives, the ^{1}H NMR spectra of which were fully in line with the assigned structures (Fig. 12). In order to assay the aminocyclitols obtained for glycosidase inhibitory activity, compounds **63b**–**65b** were transformed into their *N*-acetyl derivatives **68b**–**70b** and the free bases **68c**–**70c**, respectively, which were assayed for enzyme inhibitory activity [***I***(%)] against nine glycohydrolases: α-glucosidase (Baker's yeast), β-glucosidase (almonds), α-mannosidase (Jack beans), α-galactosidase (green coffee), β-galactosidase (bovine kidney), and α-L-fucosidase (bovine kidney), sucrase (rat small intestine), and maltase (rat small intestine). Since, as shown in conformation formulas, **68c** is related to β-L-mannopyranose-type cyclohexanepentol and **70c** is a β-D-galactopyranose analogue of valiolamine [19], they were expected to have

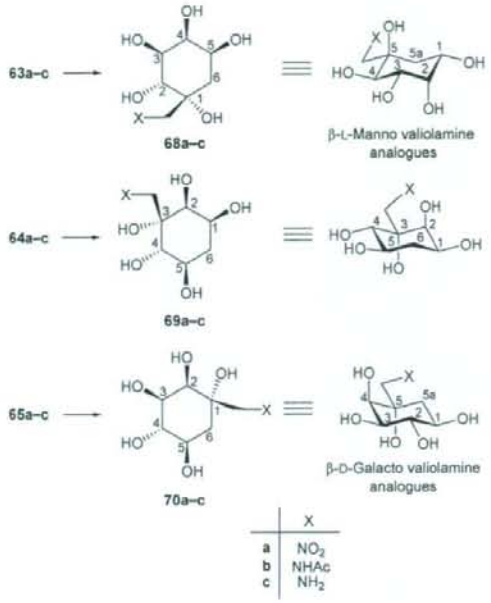

Fig. 12. Biological interesting valiolamine analogues.

some biological activity. However, only compounds **68b,c** showed inhibitory activity, very weak and limited to α-glucosidase and α-mannosidase (I = 20–30%, at 10^{-4} M).

6. Synthesis of deoxyinosamines of biological interest

The methylthio and methoxyl functions of the α-mannosidase inhibitor mannostatin A (**71a**) and **71b** [20] may match those of the 5-hydroxymethyl in the mannopyranosyl cation model (Fig. 13).

Among all 5a-carbaglycosylamines, ground-state mimicking glycosidase inhibitors, synthesized so far, 5a-carba-α-L-fucopyranosylamine (**72**) [21] was demonstrated to possess the strongest inhibitory activity against α-fucosidase. Therefore, it seemed desirable to choose it as a lead as well as mimetic compound, and new derivatives were generated by replacement of the methyl group with methyloxy elements: the methyloxy analogue, namely the 1-*O*-methyl derivative **74** of 5-amino-5-deoxy-L-*talo*-quercitol (**73**) and a series of its methyl ethers were synthesized, and their enzyme-inhibitory activity was evaluated.

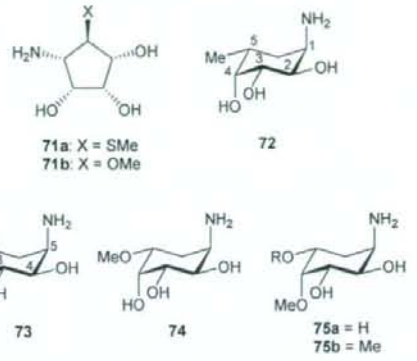

Fig. 13. 5-Amino-5-deoxy-L-*talo*-quercitol **73** and three methyl ethers **74–75a,b**, designed on the basis of the structures of α-mannosidase and fucosidase inhibitors **71a,b** and **72**.

Treatment of a mixture of di-*O*-isopropylidene derivatives (**11a** and **12a**) of (–)-*vibo*-quercitol (**5**) with SO_2Cl_2 (3 M equiv.) in the presence of DMAP in pyridine gave, after fractionation over a silica gel column, two chloro compounds **76** (40%) and **77** (58%) (Fig. 14). Azidolysis of the desired chloride **76** with NaN_3/DMF in DMF at 100°C gave the azide **78** (50%), accompanied by some elimination products. Selective *O*-deisopropylidenation of **78** was conducted under the influence of trace *p*-TsOH in MeOH. Formation of the mono-*O*-isopropylidene derivative **79** in the reaction mixture was easily monitored by use of TLC. The mixture of products was separated on a silica gel column to give **79** (71%), along with **78** (ca. 10%) and the tetrol **84** (ca. 7%). Selective tosylation of **79** was carried out by treatment with 5 M equiv. *p*-TsCl in pyridine at low temperature. When **79** just disappeared, two mono-tosylates **80** (43%) and **81** (29%), and the ditosylate **82** (15%) were produced. Compounds **80** and **81** isolated by silica gel chromatography could be readily differentiated. The structure of **80** was characterized by the ^{1}H NMR spectrum of its acetyl derivative **83**.

Thus, hydrogenolysis of **80** in ethanol containing Ac_2O in the presence of Raney nickel gave the crystalline amide tosylate **85**, quantitatively (Fig. 15). Treatment of **85** with NaOAc/90% aq MCS at

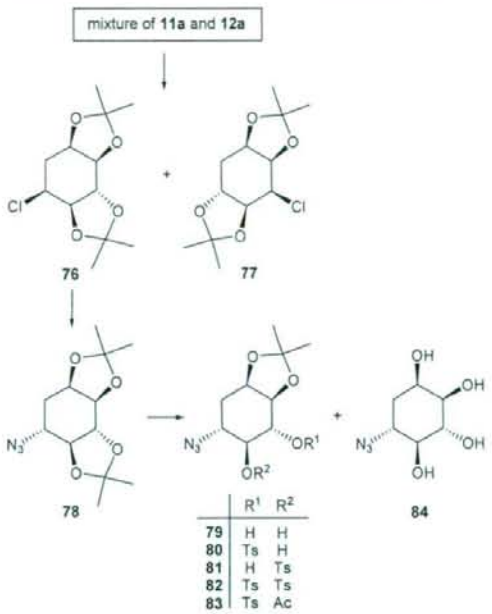

Fig. 14. Conversion of the chloride **76** into 5-azido-5-deoxy-L-*vibo*-quercitol derivatives.

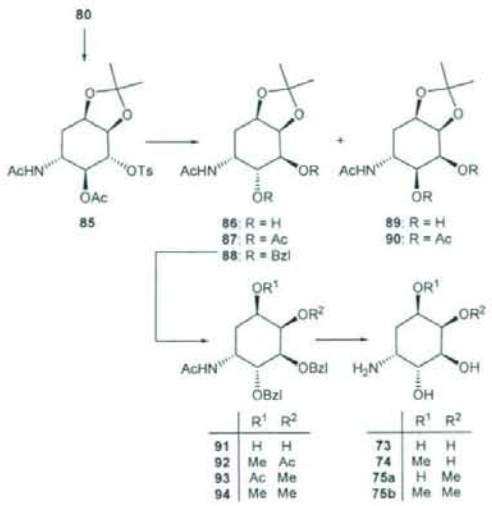

Fig. 15. Synthesis of 5-amino-5-deoxy-L-*talo*-quercitol (**73**) and some mono- and di-*O*-methyl derivatives, **74** and **75a,b**.

120°C produced an approximately 10:1 mixture of the two diols **86** and **89** with *talo*- and *allo*-configurations. Alternatively, a similar reaction was carried out in DMF to give a 1:10 mixture of **86** and **89**. On conventional acetylation the di-*O*-acetyl derivatives **87** and **90** were isolated, respectively.

Mechanistically, two compounds were likely to be produced mainly by the acetolysis of **85** (Fig. 16). Thus, in DMF, direct S_N2 reaction with an acetate ion would be undergone preferentially to afford products with an *allo*-configuration, while, on the other hand, in aq MCS the 4-acetoxyl would participate at C-3 to form an intermediate acetoxinium ion at C-3 and 4, which would be cleaved to give rise to two products with *allo*- and

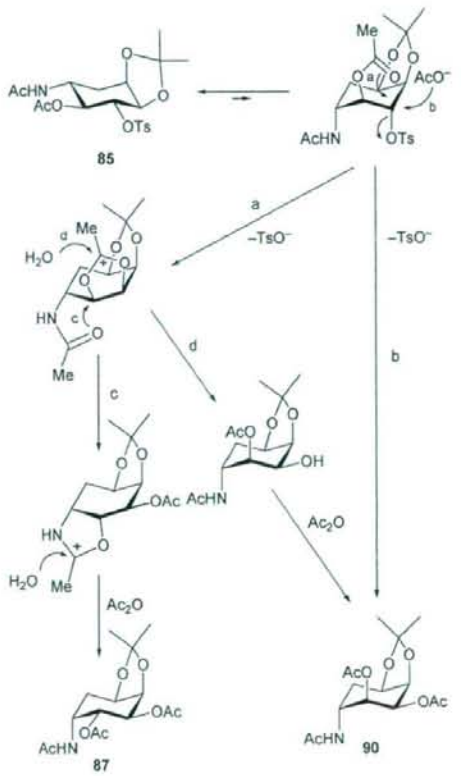

Fig. 16. Postulated reaction mechanism for formation of compounds **86** and **90**.